1. GREAT Summary

Introduction

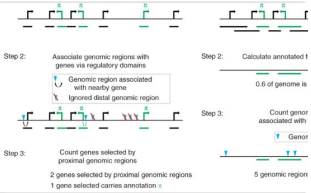
GREAT: Genomic Regions Enrichment of Annotations Tool

- Online software

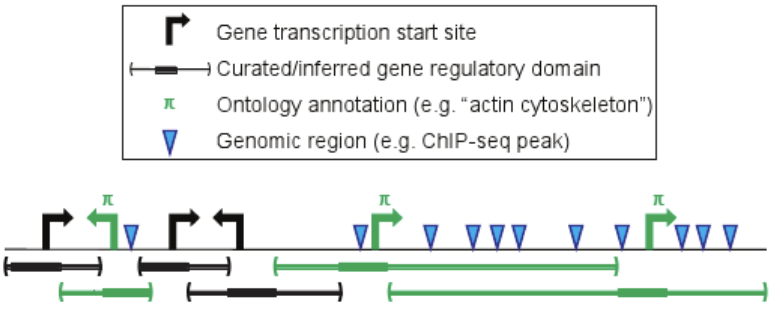
> GREAT Input: Genomic Regions Enrichment of Annotations Tool, Bejerano Lab, Stanford University
>
> News Aug. 7, 2020: This issue is now fixed! There was a server issue in GREAT partly due to an increasing service demand. We are trying to fix the issue. In the meanwhile, if you are pushing a large number of jobs, please increase the delay between jobs.
>
> http://great.stanford.edu/public/html/

- Paper

> GREAT improves functional interpretation of cis-regulatory regions - Nature Biotechnology
>
> ChIP-Seq data are usually analyzed with approaches developed for microarrays, which only consider binding events within a few kilobases of a gene. McLean et al. present an algorithm that takes into account more distant events, thereby improving functional annotation of regulatory regions.
>
> https://www.nature.com/articles/nbt.1630



- Function: model cis-regulatory regions through the use of long-range regulatory domains and a genomic region-based enrichment test.



Binomial test over genomic regions
n = 11 total genomic regions
p_π = 0.75 fraction of genome annotated with π
k_π = 11 genomic regions annotated with π
P-value = 0.04

Hypergeometric test over genes
N = 6 total genes
K_π = 3 genes annotated with π
n = 3 genes with an associated genomic region
k_π = 3 genes annotated and with a genomic region
P-value = 0.05

- Long-range regulatory domains: consisting of a basal domain that extends 5kb upstream and 1kb downstream from its transcription start site, and an extension up tp the basal regulatory domain of the nearest upstream and downstream genes within 1Mb. (marked by the horizontal line in the figure above)

  rules and distances could be modified by users

- Genomic region-based enrichment test: binomial test over genomic regions and hypergeometric test over genes
  - The binomial test calculates the P value of the observed enrichment for term π as the probability of selecting annotation π at least k times in n attempts
    1. n is the total number of genomic regions in the input set.
    2. p_π is the a priori probability of selecting a base pair annotated with π when selecting a single base pair uniformly from all non–assembly gap base pairs in the genome.
    3. k_π is the number of genomic regions in the input set that cause annotation π to be selected.

$$\sum_{i=k_\pi}^{n} \binom{n}{i} p_\pi^i (1-p_\pi)^{n-i}$$

  - The hypergeometric test calculates the P value of the observed enrichment for term π as the fraction of ways to choose n genes without replacement from the entire group of N genes such that at least k_π of the n possess ontology annotation π
    1. N is the total number of genes in the genome.
    2. K_π is the number of genes in the genome that possess ontology annotation π.
    3. n is the number of genes selected because one or more input genomic regions resides in their regulatory domains.
    4. k_π is the number of selected genes that possess ontology annotation π.

$$\sum_{i=k_\pi}^{\min(n,K_\pi)} \frac{\binom{K_\pi}{i}\binom{N-K_\pi}{n-i}}{\binom{N}{n}}$$

Input & Parameters

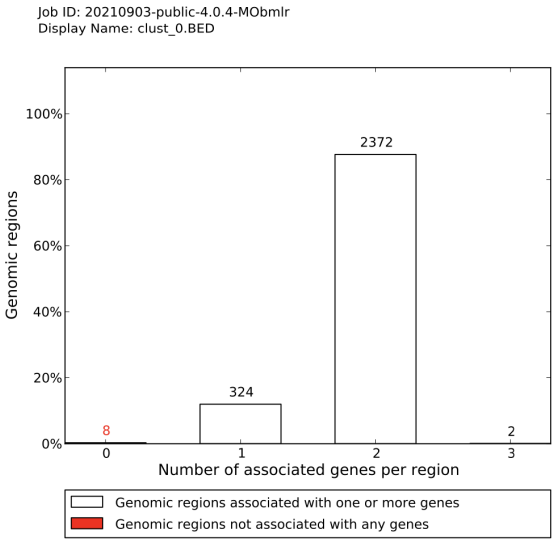
1. Input: 10 clusters from ATAC(PBMC)
2. Parameters:
   - Species Assembly: Human: GRCh38
   - others are default

Output

cluster_0, for example:

1. Region-Gene Association Graphs

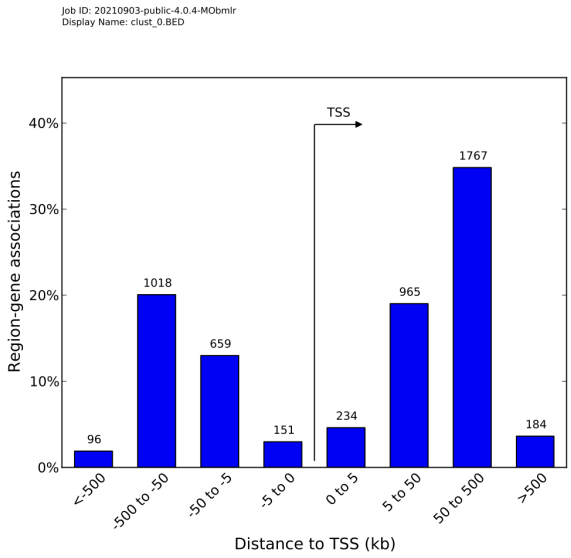
- **Number of associated genes per region:** how many genes each input genomic region is assigned as putatively regulating based on the association rule used.



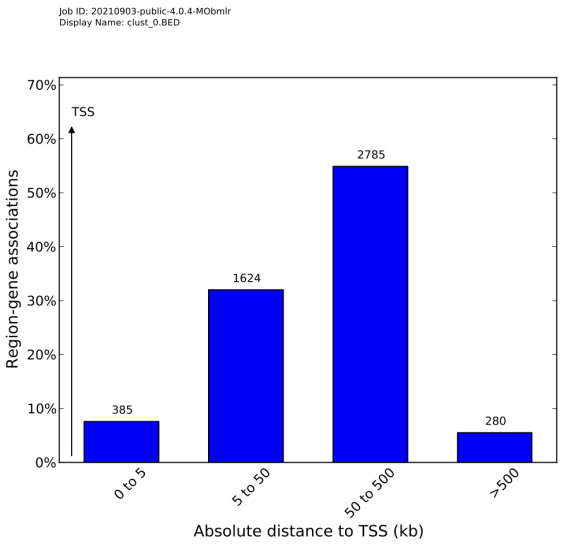
- **Binned by orientation and distance to TSS:** show the distance between input regions and their putatively regulated genes.

  TSS: transcription start site

  One input region may be associated with 2 or 3 genomic regions(see figure of number of associated genes per region).



- **Binned by absolute distance to TSS:** only the distance to TSS is considered.



2. Global export

"xxx.csv" contains GO ontology results based on the binomial test and the hypergeometric test.

GREAT ranks results by the **binomial p-value**, and it consider this the single best way to examine genome-wide cis-regulatory datasets. It accounts for biases in gene regulatory domain size and provides an accurate picture of the cis-regulatory landscape.

# GREAT version 4.0.4	**Species assembly: hg38**	**Association rule: Basal+extension: 5000 bp upstream, 1000 bp downstream, 1000000 bp max extension, curated regulatory domains included**											
# Ontology	Term Name	Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
GO Biological Process	regulation of immune system process	1	1.520726E-76	2.0011233434E-72	2.119458	651	0.2405765	8	2.375372211875E-08	1.474248	263	1423	0.1117247
GO Biological Process	cell activation	3	9.487898E-76	4.161708326066667E	2.364982	528	0.195122	55	0.000076582485020(	1.425243	181	1013	0.0768904
GO Biological Process	leukocyte activation	4	1.189875E-75	3.91439128125E-72	2.529586	471	0.1740576	52	0.000036690404609(	1.477835	161	869	0.06839422
GO Biological Process	lymphocyte activation	5	6.550268E-67	1.72389953224E-63	3.177439	299	0.1104952	2	2.0458849978E-08	2.085844	91	348	0.0386576
GO Biological Process	positive regulation of immune system process	6	1.127757E-62	2.4733590605E-59	2.300966	466	0.1722099	6	2.7404736015E-08	1.618448	182	897	0.07731521

Ontologies and columns are below

Some clusters don't have three domains of GO. Some clusters have human phenotype results.

Ontologies

Show Non-Empty | Expand All | Collapse All

GO

- [x] GO Molecular Function
- [x] GO Biological Process
- [x] GO Cellular Component

Phenotype

- [ ] Mouse Phenotype
- [ ] Mouse Phenotype Single KO
- [ ] Human Phenotype

Genes

- [ ] Ensembl Genes

Click a link above to jump to the corresponding table.

What data does each ontology provide?

Can I use other ontologies?

Columns

Show All | Hide All

General

- [ ] Ontology Term ID

Binomial over Regions

- [x] Rank
- [x] Raw P-Value
- [ ] Bonferroni P-Value
- [x] *FDR Q-Value*
- [x] Fold Enrichment (Obs/Exp)
- [ ] Expected (n * p)
- [x] Observed Region Hits (k)
- [ ] Genome Fraction (p)
- [x] Region Set Coverage (k/n)

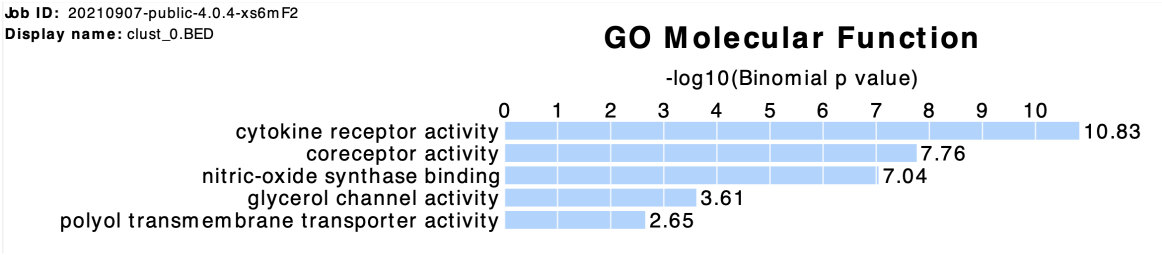
Hypergeometric over Genes

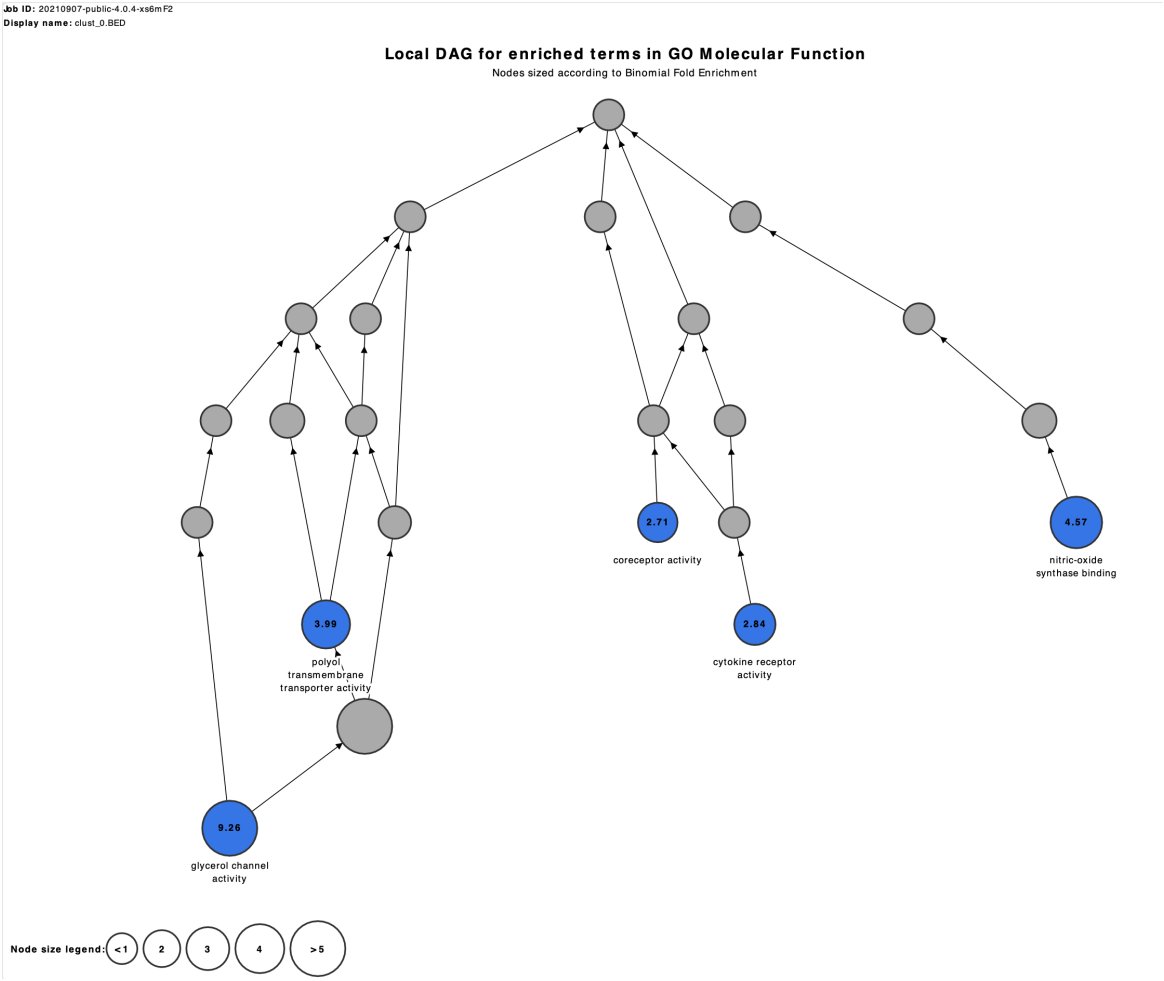
- [x] Rank
- [ ] Raw P-Value
- [ ] Bonferroni P-Value
- [x] *FDR Q-Value*
- [x] Fold Enrichment (Obs/Exp)
- [ ] Expected (n * K/N)
- [x] Observed Gene Hits (k)
- [x] Total Genes (K)
- [x] Gene Set Coverage (k/n)
- [ ] Term Gene Coverage (k/K)

3. GO enrichment visualization

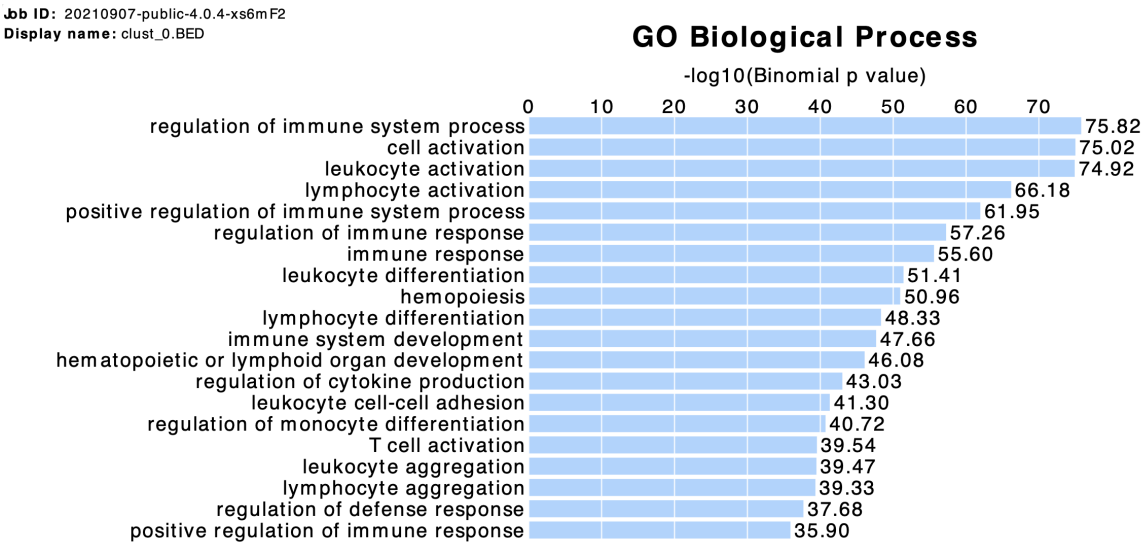
GREAT provides bar chart and DAG.

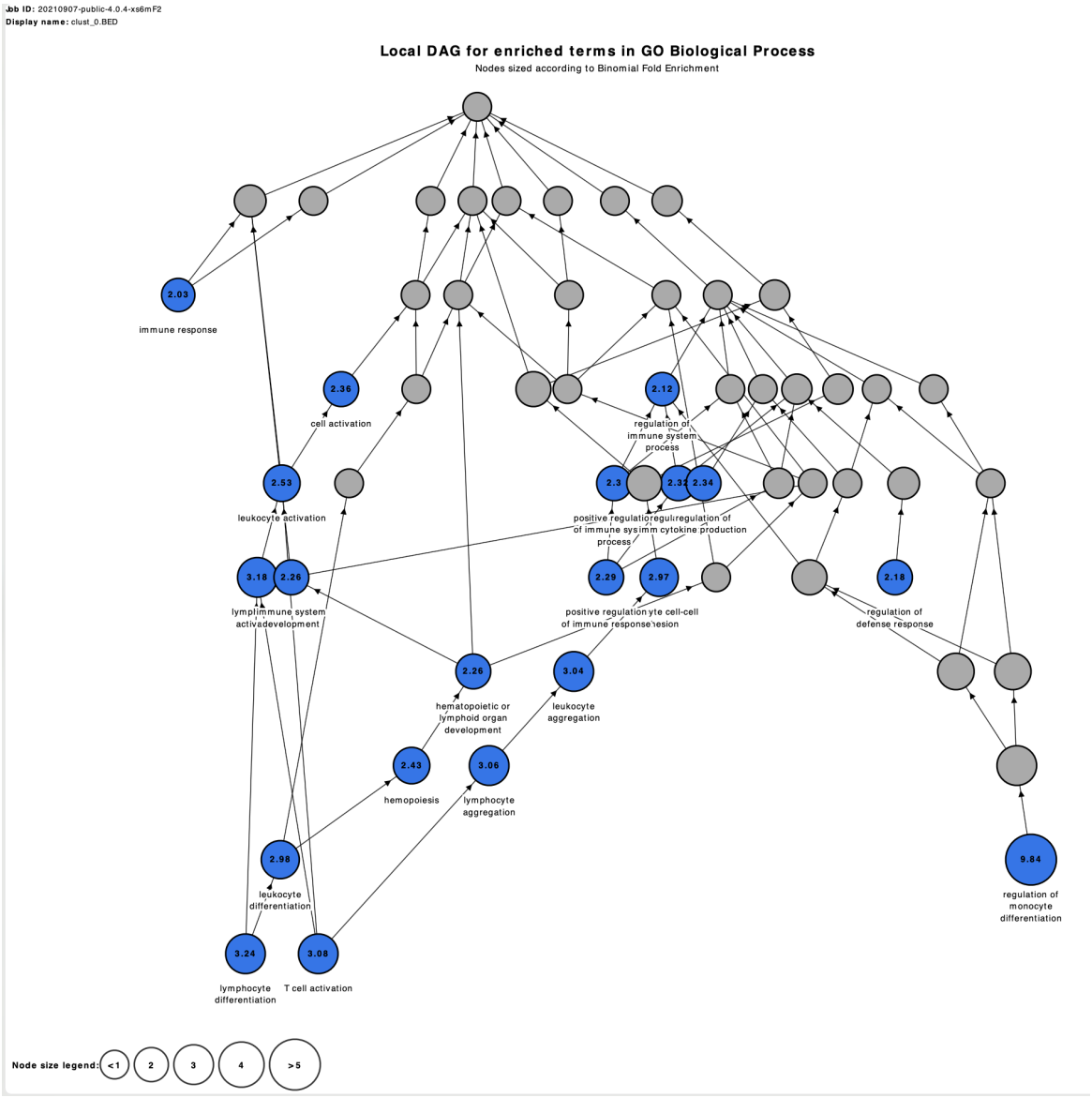
- GO molecular function



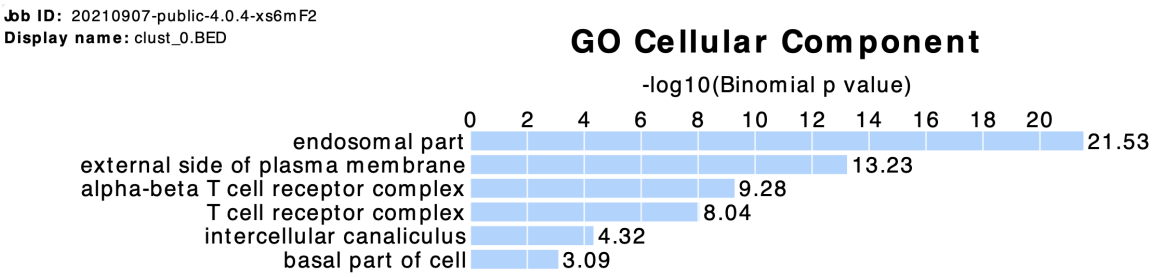


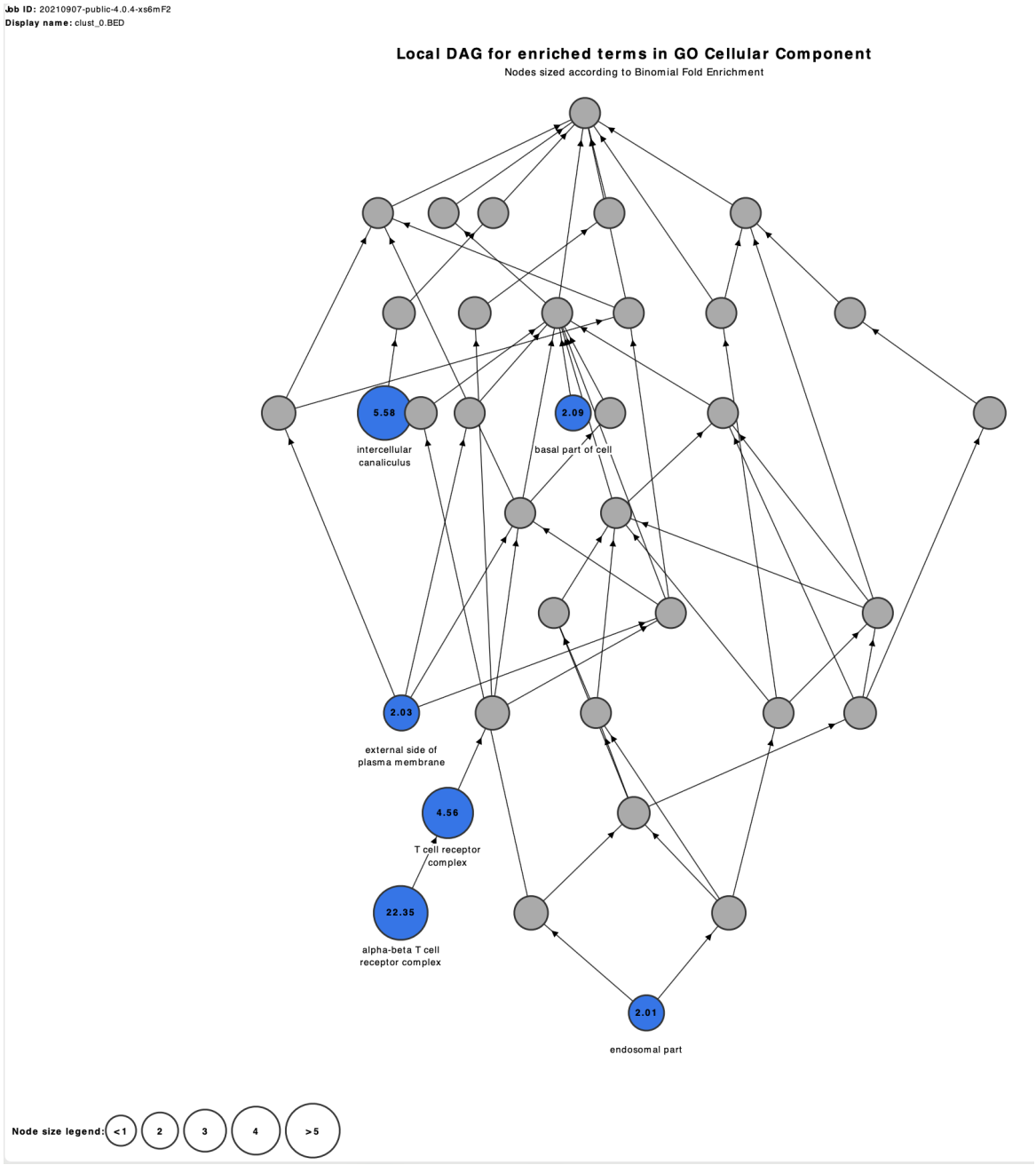
- GO biological process





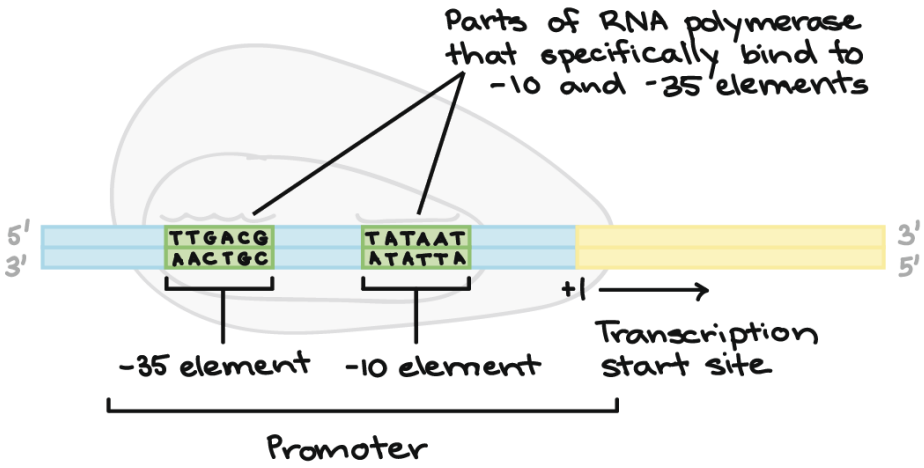
- GO cellular component



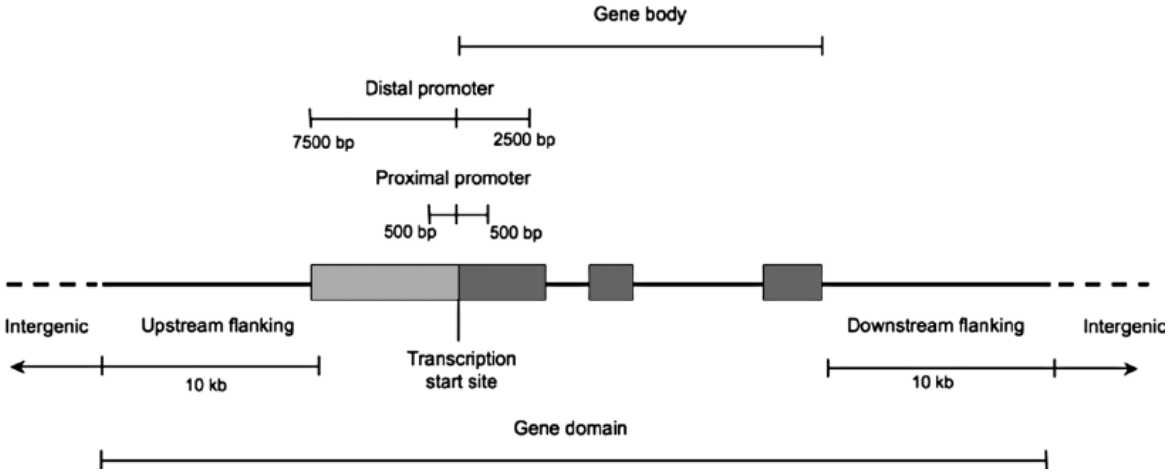


Biological terminology

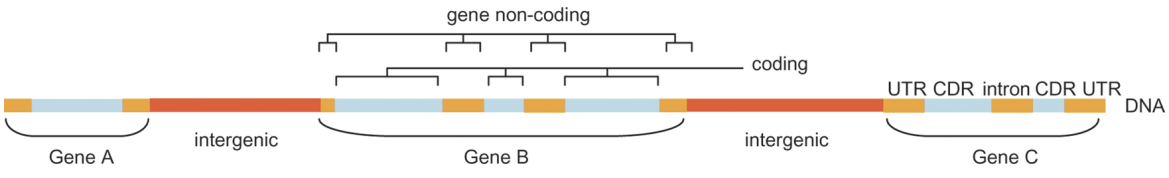
- **TSS: transcription start site**



- **Genomic region:** contains a gene body with exons (dark bars) and introns, and a promoter



- **Intergenic region**

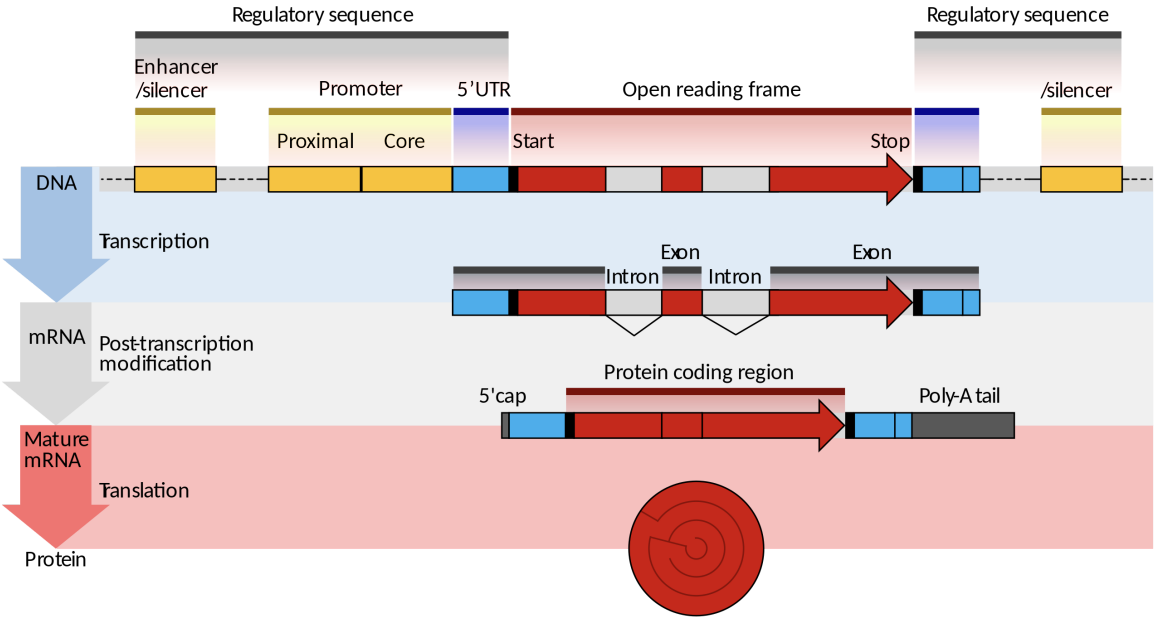


UTR: Untranslated region; CDR: Coding region

- **Flanking region: 5′ flanking region and 3' flanking region**

  5′ flanking region: a region of DNA that is adjacent to the 5′ end of the gene. The 5′ flanking region contains the promoter, and may contain enhancers or other protein binding sites.

  3′ flanking region: The region of DNA which borders the 3' end of a transcription unit and where a variety of regulatory sequences are located.


Regulatory sequence
Regulatory sequence
Enhancer
/silencer
Promoter
5'UTR
Open reading frame
/silencer
Proximal
Core
Start
Stop
DNA
Transcription
Exon
Exon
Intron
Intron
mRNA
Post-transcription
modification
Protein coding region
5'cap
Poly-A tail
Mature
mRNA
Translation
Protein